Lab 1 Question 1: Age and Voting Patterns

Max Hoff, Connor McCormick, and Chi Ma

Contents

1 Are Democractic voters older or younger than Republican voters in 2020?

1.1 Importance and Context

The 2020 presidential election has been talked about as a potential turning point in American politics. After much stability in terms of demographic voting patterns over the previous 40 or so years, starting in the mid-2010's, signs of stability for these voting patterns began to weaken. As these voting patterns changed, so have the policies that each party endorses. Pundits began speculating about the future of each party, and in turn the direction of the country. Understanding the underlying changes in demographics, and understanding how those demographics relates to voting patterns is potentially insightful for predicting how the U.S. will determine it's future.

This study focuses on one of the main demographic attributes; age. Republicans have historically been an older voting group, but with all the shifting party alliances, is this trend still holding? If it is changing, this might have a large impact on future elections, and which policies each party embraces.

Our study concludes that, most likely, this trend is, in fact, still holding. There are some uncertainties in our study, so in the rest of this paper we will detail how we reached these conclusions, giving you a better understanding of how reliable these results are and hopefully improving your understanding of the current political climate.

1.2 Exploratory Data Analysis

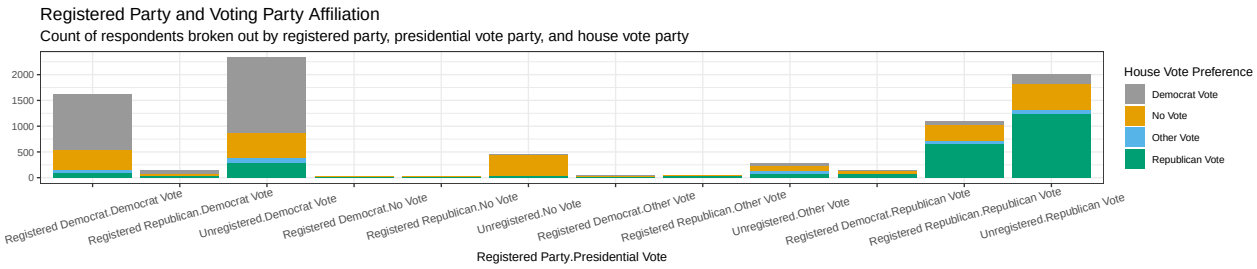
We reached this conclusion by analyzing the American National Election Survey (ANES) data set, which surveys thousands of American citizens on their political beliefs. The data set includes many fields related to party affiliation and age. The main challenge of our analysis was determining how we defined 'party', and then how to conduct a test.

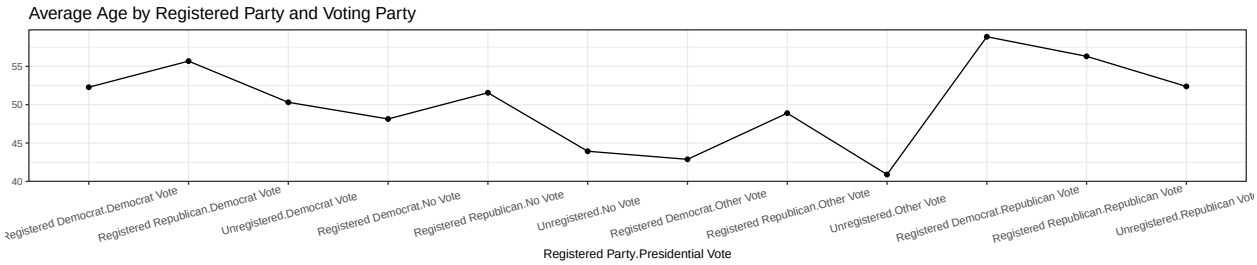
1.2.1 Determining Party Affiliation

Although the question seems fairly straightforward at the outset, once you get into the details, the assignment of an individual to a political party is actually quite tricky. Is a moderate voter that consistently votes Republican a Republican? Is a registered Democrat that votes for a Republican presidential candidate, an independent candidate in the Senate, and a Democrat in the House a Democrat? What if these cases make up a significant portion of the population, and have a material age difference compared to the rest? Let's dig into some of these cases and uncover a grouping that makes the most sense for this test.

Our main concern with picking presidential vote as a respondent's party is that the 2020 election was an unusual election in which we believe a material proportion of the population deviated from their typical voting party. Our main concern with choosing registered party as a respondent's party is that all voters might not be registered, and that the age of those registered might be skewed from the rest of the population.

The first question that might come to mind is; 'Is party identification closely tied to voting party?'. The chart below shows an unexpected picture; that declared party and voting party aren't as correlated as we were expecting:





A couple observations from the charts above:

1. Unregistered voters outnumbered registered voters for both Democrats and Republicans(!).
2. From the line chart you can see that age can have significant differences between groups, meaning the grouping decision will most likely have a material impact on the test.
3. Most voters registered to a party vote (shown by the close to zero amount for 'Registered Democrat.No Vote' and 'Registered Republican.No Vote').
4. There's a material amount of Republicans that voted Democrat (and vice versa).
5. For unregistered voters, a material amount of respondents voted for conflicting parties between the Presidential vote and the House vote.

Through this analysis we were hoping to find that either voting party or registered party gave us a clean grouping, but unfortunately this didn't happen, as shown by the messiness of the observations above. With that being said, we think it's best to go with the grouping that makes the most logical sense. To us, we believe registered party makes the most sense since (1) voting party for this year is most likely flawed, (2) there is a large portion of conflicting house and presidential votes and (3) that our sample size of only registered voters is large enough. Although we could come up with a grouping that combines registered voters with voting party for unregistered voters, we ran into the problem of having to decide whether to include or exclude respondents that support the opposing party with their House or Senate vote. We believe the cleanest cut-off is to go with registered party. Although this decision is subjective in nature and will impact the results of our test, there was no clear-cut option, and each team member believed this to be the least-worst choice.

1.2.2 Age Values

The given age field was much simpler compared to determining party. The only item to note is that roughly 5% of respondents did not give an age. We do not believe this will negatively impact our test results too much though given that the respondents that did not give an age had a similar voting distribution compared to respondents that gave an age (charts withheld for brevity's sake). In other words, we don't believe the non-respondents will skew our results much.

1.3 Hypothesis

Our team believes that Republicans voters have historically been older than Democrat voters and this trend most likely held in the 2020 election. The way we'll conduct our test though is that we'll set out to prove or disprove that Republican and Democrat voters have the same age, and then determine direction if the test shows that the average ages are in fact different.

In terms of statistical language, we'll conduct an unpaired two-tailed t-test, against the null hypothesis that Democrat and Republican voters have the same age on average. Our alternative hypothesis is that there is a difference in average age between the two groups.

The reason we chose a two-tailed unpaired t-test is because of the met assumptions in the underlying data. The data is unpaired since a voter cannot prefer both a Democrat and a Republican. The metric we are evaluating, age, is a metric variable. The data is fairly close to normally distributed and our data is mostly independent and identically distributed. There is a possibility that each respondent isn't fully independent of

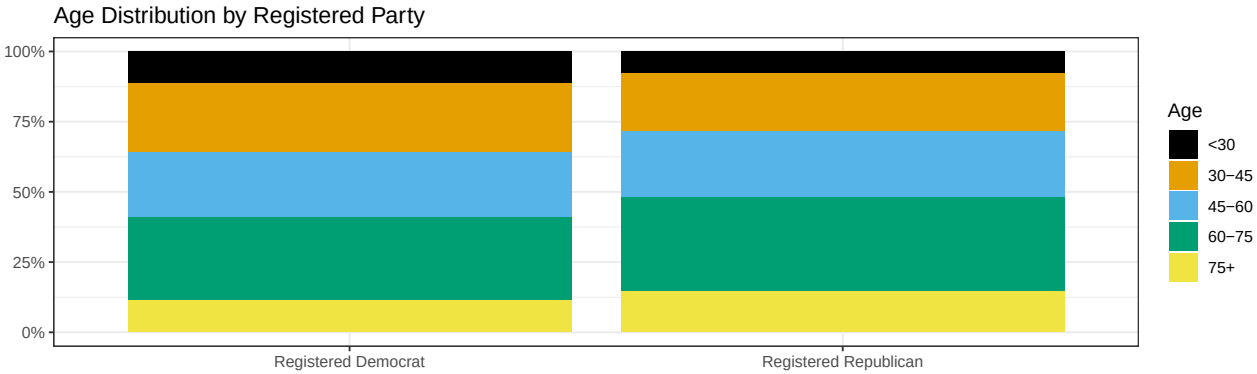
other respondents since one respondent could tell others similar to them about it, but we suspect that this effect is small enough to not invalidate our findings.

1.4 Test Results

```
test_result <- df %>%
  filter(party_pre_preference_simplified == 'Registered Democrat' | party_pre_preference_simplified ==
  t.test(age ~ party_pre_preference_simplified)
```

The resulting p-value of 0.027 is less than 0.05, which in standard practice tells us to reject the null hypothesis. In plain language, this means that the difference in ages between the two groups is most likely real. There's a 1-in-40 that the result isn't real (a false positive/Type-1 error), which, in our opinion, is fairly reliable. A 1-in-40 chance is not zero though, so these results do need to be taken with a grain of salt.

Now that we've determined that there is a difference, we need to decide which direction this difference is in. At first glance, it looks like Republicans are in fact older voters, given they have a higher average of 52 versus 51 for Democrats. To confirm this though, let's take a look at the distribution for each of the parties:



As you can see, the age distribution for each party shows a similar picture; Republicans are older than Democrats.

An additional approach to evaluating our results is to analyze the effect size by using a method called "Cohen's D". This method outputs a single numerical value that tells you how different the distributions are. Our result of 0.08 tells us that this effect is very small (a small effect size is around 0.2, and a large effect size is around 0.8), which ties with how close the distribution of ages are in the chart above.

Although our results seem fairly reliable, there is still quite a bit of work that can be done to confirm our findings. We only used one data source in this study, so it would be interesting to see how our results compare to a study conducted on a entirely separate data set. On top of this, although we answered our original question of which party contained older voters, it would be interesting to look deeper into the degree of this difference and how this difference has changed over time. From our findings it seems like this difference was fairly small (roughly a year's difference in average age), but more nuanced understandings of these differences could yield impactful findings, and potentially improve predictions about the future of the U.S.